Published online 17 November 2010 Nucleic Acids Research, 2011, Vol. 39, Database issue D245–D252
doi:10.1093/nar/gkq940

CLIPZ: a database and analysis environment for experimentally determined binding sites of RNA-binding proteins

Mohsen Khorshid, Christoph Rodak and Mihaela Zavolan*

Biozentrum, University of Basel and Swiss Institute of Bioinformatics, Klingelbergstrasse 50-70, 4056 Basel, Switzerland

Received August 13, 2010; Revised September 29, 2010; Accepted September 30, 2010

ABSTRACT

The stability, localization and translation rate of mRNAs are regulated by a multitude of RNA-binding proteins (RBPs) that find their targets directly or with the help of guide RNAs. Among the experimental methods for mapping RBP binding sites, cross-linking and immunoprecipitation (CLIP) coupled with deep sequencing provides transcriptome-wide coverage as well as high resolution. However, partly due to their vast volume, the data that were so far generated in CLIP experiments have not been put in a form that enables fast and interactive exploration of binding sites. To address this need, we have developed the CLIPZ database and analysis environment. Binding site data for RBPs such as Argonaute 1-4, Insulin-like growth factor II mRNA-binding protein 1-3, TNRC6 proteins A-C, Pumilio 2, Quaking and Polypyrimidine tract binding protein can be visualized at the level of the genome and of individual transcripts. Individual users can upload their own sequence data sets while being able to limit the access to these data to specific users, and analyses of the public and private data sets can be performed interactively. CLIPZ, available at http://www.clipz.unibas.ch, aims to provide an open access repository of information for post-transcriptional regulatory elements.

INTRODUCTION

Almost all cellular RNAs interact with RNA-binding proteins (RBPs) to form ribonucleoprotein complexes (RNPs). The overall composition and precise architecture of these RNPs undergo dynamic remodeling in response to signals and cellular state. Initial annotation (1) indicated that the human genome contains ∼300 genes that encode proteins with an RNA-recognition motif (RRM). This is only one of the over 40 distinct protein domains known to contact RNA. RBP–RNA interactions are highly context dependent and many RBPs carry out different functions in different cellular compartments. For instance, the T-cell intracellular antigen 1 (TIA-1) functions as a splicing factor in the nucleus; it binds to an intronic splice enhancer in the Fas pre-mRNA leading to the inclusion of the proximal exon (2). In the cytoplasm, TIA-1 regulates the stability of mature mRNAs: its binding to AU-rich elements located in the 3′ untranslated regions (3′UTRs) of mRNAs (such as that of transforming growth factor beta, TGFβ) attracts the mRNA degradation machinery. The same AU-rich element in the TGFβ 3′UTR when bound by the HuR RBP leads to mRNA stabilization (2). Thus, precise knowledge of spatio-temporal associations between RBPs and mRNAs under various conditions is key to understanding how the level, translation rate and cellular localization of those mRNAs are regulated during the life time of a cell.

With some exceptions, such as the knowledge-based potential function designed by Zheng *et al.* (3) to predict the specificity and relative binding energy of RNA-binding proteins, computational models describing the binding specificity of RBPs (by contrast, for instance, with transcription factors) are lacking (4). Recently, however, experimental methods for high-throughput and high-resolution identification of RBP binding sites have been developed. They rely on cross-linking and immunoprecipitation (CLIP) of RBPs of interest (5) followed by deep sequencing (6–8). In a particular variant of CLIP, termed PAR-CLIP (photoactivatable-ribonucleoside-enhanced cross-linking and immunoprecipitation), the incorporation of photo-reactive nucleotides in mRNAs prior to cross-linking induces a specific mutational signature in the sequenced reads relative to the reference

*To whom correspondence should be addressed. Tel: +41 61 267 15 77; Fax: +41 61 267 15 84; Email: mihaela.zavolan@unibas.ch genome, thereby enabling the separation of cross-linked binding sites from other RNA fragments that are captured non-specifically during the experiment (7). Many questions concerning the function, specificity and modulation of activity of RBPs can be addressed through analyses of corresponding PAR-CLIP data sets. For example, the sites with the highest number of cross-linking events (indicated by T-to-C mutations in the sequenced reads) can be analyzed to uncover the sequence specificity of the RBP and to identify cellular pathways that are targeted by the RBP in a concerted manner. Moreover, with PAR-CLIP data available for multiple RBPs, one can begin to identify regions of cross-talks between multiple RBPs on individual mRNAs.

Here, we describe a database of binding sites that we constructed based on CLIP data for various proteins that are known to regulate mRNA splicing (polypyrimidine tract binding protein), stability and/or translation rate (Quaking, Pumilio2, Argonautes 1-4, TNRC6 A-C, Insulin-like growth factor II mRNA-binding proteins 1–3). The data are presented through a web interface that supports not only visualizations but also further analyses of RBP binding sites. The platform also allows registered users to submit for functional annotation short reads resulting from CLIP, small RNA sequencing and mRNA sequencing experiments. Once uploaded, these data can be explored through various interactive analysis tools that we developed. Due to its user- and dataset-management system, the platform can support collaborative projects involving private and public data and multiple users. This resource is of great value to researchers that study the mechanisms regulating mRNA stability and translation.

MATERIALS AND METHODS

Sequence annotation

The computational pipeline underlying the construction of the CLIPZ database takes as input fasta-formatted files of sequences that were obtained from CLIP samples through deep sequencing. These sequences are submitted to an initial annotation process that attempts to identify the origin (within the genome and within known transcripts) of individual sequence reads. The annotation procedure is described in detail elsewhere (9). Briefly, it consists of adaptor removal, mapping of sequenced reads to the genome and to known transcripts and functional annotation of each read based on its best mappings. A sketch of the data flow is shown in Figure 1.

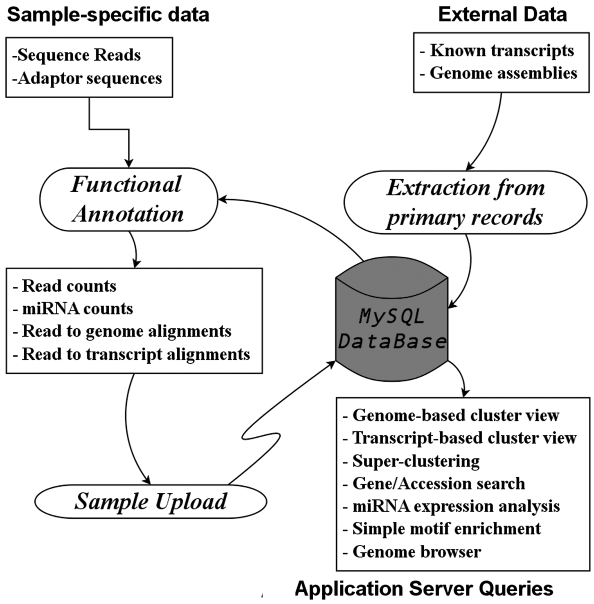


Figure 1. CLIPZ Data flow. Procedures are further described in the 'Materials and Methods' section.

Adaptor removal. During sample preparation, adaptors are ligated at both 5′ and 3′ ends of CLIP sequence fragments. Because most of the CLIP data that is currently available has been generated using the Solexa sequencing technology (10), our procedure for adaptor removal is specific to this technology (though other adaptor configurations can easily be taken into account). The 5′ adaptor serves as a sequencing primer, and we expect that only the 3′ adaptor (or part of it) is sequenced. We use an in-house ends-free local alignment algorithm (11) (parameters: 2 for match, −3 for mismatch, −5 for gap opening, −2 for gap extension) to align the 3′ adaptor to the reads. The part of the sequence read that aligns to the 5′ end of the 3′ adaptor is removed, and if the remainder of the read is longer than 15 nt, it is retained for further analysis. Distinct sequences are deposited in the database together with their copy number in the sample under study.

Sequence mapping. All distinct sequences are mapped to the genome assembly. Currently, the database contains CLIP samples obtained from human cells, for which we used the *hg18* version of the human genome assembly from the University of California at Santa Cruz (http://genome.cse.ucsc.edu), but analyses of mouse data sets are also supported. Because not all transcripts that have been sequenced and are present in sequence databases can be mapped to the genome assembly and because various contaminants can be found in CLIP samples, we also map the reads to a database of sequences with known function (ribosomal, transfer, small cytoplasmic, small nuclear and small nucleolar RNAs, PIWI proteins-associated RNAs, miRNAs, messenger RNAs, miscellaneous non-coding RNAs obtained from sequencing projects, bacterial and fungal ribosomal RNAs, genomes of common bacteria, vector, adaptor and size marker sequences). The sources of these sequences are as follows.

- **Protein-coding as well as non-coding sequences** (mRNA, tRNA, scRNA, snoRNA, rRNA, snRNA, piRNA and miscRNA) were extracted based on the 'molecule type' and 'feature key' fields of the Genbank records (Genbank release of 19 January 2010).
- **Bacterial and fungal sequences** (ribosomal RNA and complete genome sequences of bacterial and fungal species available in the Nucleotide Database of Genbank) were included in order to detect contaminations.

- **Vector sequences** from ftp://ftp.ncbi.nih.gov/pub/UniVec/UniVec were included for the same purpose.
- **miRNAs precursor sequences** from ftp://mirbase.org/pub/mirbase/CURRENT were included to be able to analyze data specific to proteins involved in the miRNA-dependent silencing pathway.
- **Miscellaneous non-coding RNAs** from http://www.noncode.org were included to enable detection of interactions involving poorly characterized non-coding RNAs, that have not been categorized yet in Genbank databases.
- **Human tRNAs** from http://lowelab.ucsc.edu/GtRNAdb/Hsapi/Hg17-tRNAs.fa and
- **Mouse tRNAs** from http://lowelab.ucsc.edu/GtRNAdb/Mmusc/Mm6-tRNAs.fa were included because these resources provide extensive annotations of tRNAs.
- **Repeat masker annotations** of the genome provided by http://genome.cse.ucsc.edu were used to detect repeat elements.

To align sequence reads to target sequences, we use the 'Oligomap' algorithm (9) that exhaustively reports the mappings with 0 or 1 error (mismatch, insertion or deletion). The 'Oligomap' software can be downloaded from http://www.mirz.unibas.ch/software.php. In principle, we take into account all the possible loci for a given sequence read and we assume that the read originated from any of these loci with equal probability. Based on the *GMAP* (12) mappings of mRNAs to genome, we determine whether a genome-mapped read falls inside an intronic or exonic region. Based on the coding region annotation of transcripts in Genbank, we determine whether the exonic sequence reads originate from the 5′UTR, CDS or 3′UTR region of the individual transcripts. Sequence reads that map to regions with alternative splicing are assigned fractional numbers that denote the proportion of transcripts in which the region appeared in a particular section of the transcript.

Sequence annotation. Whenever an extracted sequence read maps to one or more known sequence(s) of the same functional category, that functional category is readily transferred to the sequence read. There are, however, sequence reads that map equally well to known sequences of different functional categories (e.g. tRNA, rRNA, mRNA and repeat). In these cases, we assign a functional annotation with the following priority scheme rRNA > tRNA > snRNA > snoRNA > scRNA > miRNA > piRNA > repeat > miscRNA > mRNA (reflecting roughly the abundance of various types of sequences in the cell).

Generation of clusters of sequence reads

Initial analysis of PAR-CLIP data indicated that the sequence reads obtained in individual experiments generally form well-delimited, relatively short (20–40 nt) clusters. When the binding specificity of the protein was already known, the clusters obtained from PAR-CLIP data typically *contained* the sequence motif known to represent the binding site of the protein (7). We therefore use a cluster as the central unit for data analysis and visualization. Two sequences are placed in the same cluster if they overlap by at least one nucleotide in their genomic or transcript location. We note that in data sets obtained with other CLIP protocols, the correspondence between clusters that are generated this way and individual RBP binding sites may not be as clear as it is in PAR-CLIP. As more data generated with different variants of CLIP becomes available, the definition of the visualization unit (ideally the RBP binding site) may need to be revised accordingly. Furthermore, in PAR-CLIP experiments T-to-C mutations are indicative of cross-linked positions and our analysis has shown that clusters with the largest number of T-to-C mutations are most enriched in functional binding sites for the studied RBP. The number of T-to-C mutations as well as other statistics are therefore computed for each cluster and made available in the interface. The user can sort the clusters based on these computed features in order to extract the targets that are most frequently bound by the RBP of interest.

Data storage

We use a *MySQL 5* database management system to store the results of the functional annotation process and to support downstream analyses. The database contains the following types of tables:

- **User-management-related:** tables that store information about the user (name, the group in which he/she is a member, the host laboratory, etc.).
- **Known-sequence-related:** tables that contain information about transcripts of known function that we obtain from external sources and used for short read annotation (the sequences themselves, genomic loci, NCBI Entrez Gene information when available, etc.).
- **Sample-data-related:** tables that contain information about the sequences from a submitted sample (e.g. extracted sequence reads, genomic loci, mapping coordinates within transcripts of known function, etc.).
- **Sequence-read-cluster-related:** tables that contain information about clusters of overlapping sequences typically representing individual binding sites. The cluster information is used in various visualizations.

In order to maximize the efficiency of processing subsequent queries, database tables are generated for each individual sample (for the detailed description of the database schema see the 'Help' pages provided on the web site).

Analysis environment

The software supporting the web-based queries has the following components (see Figure 2).

Web Server. The web server is responsible for the validation of the user inputs and for rendering the results of various computations. It uses *PHP 5* and a *Model View Controller (MVC)-Framework* that we developed. It communicates with the application server using a freely available PHP-Java bridge from http://php-java-bridge.sourceforge.net/pjb/.

Application Server. The application server, implemented in *Java 1.6*, provides functions that can be accessed by the web server, such as applying the functional annotation pipeline to an uploaded sample. It is also responsible for process control, logging the job outputs and reporting the errors whenever jobs fail. Due to the large volume of typical CLIP data sets, we employ a *PC-Cluster* for parallel processing. The job distribution to the cluster and the handling of conflicts that may result from multiple parallel-running jobs requesting the same data/resource at the same time are also handled by the application server.

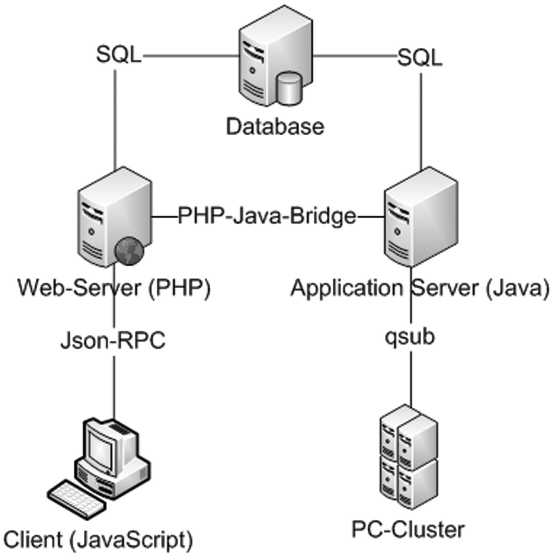
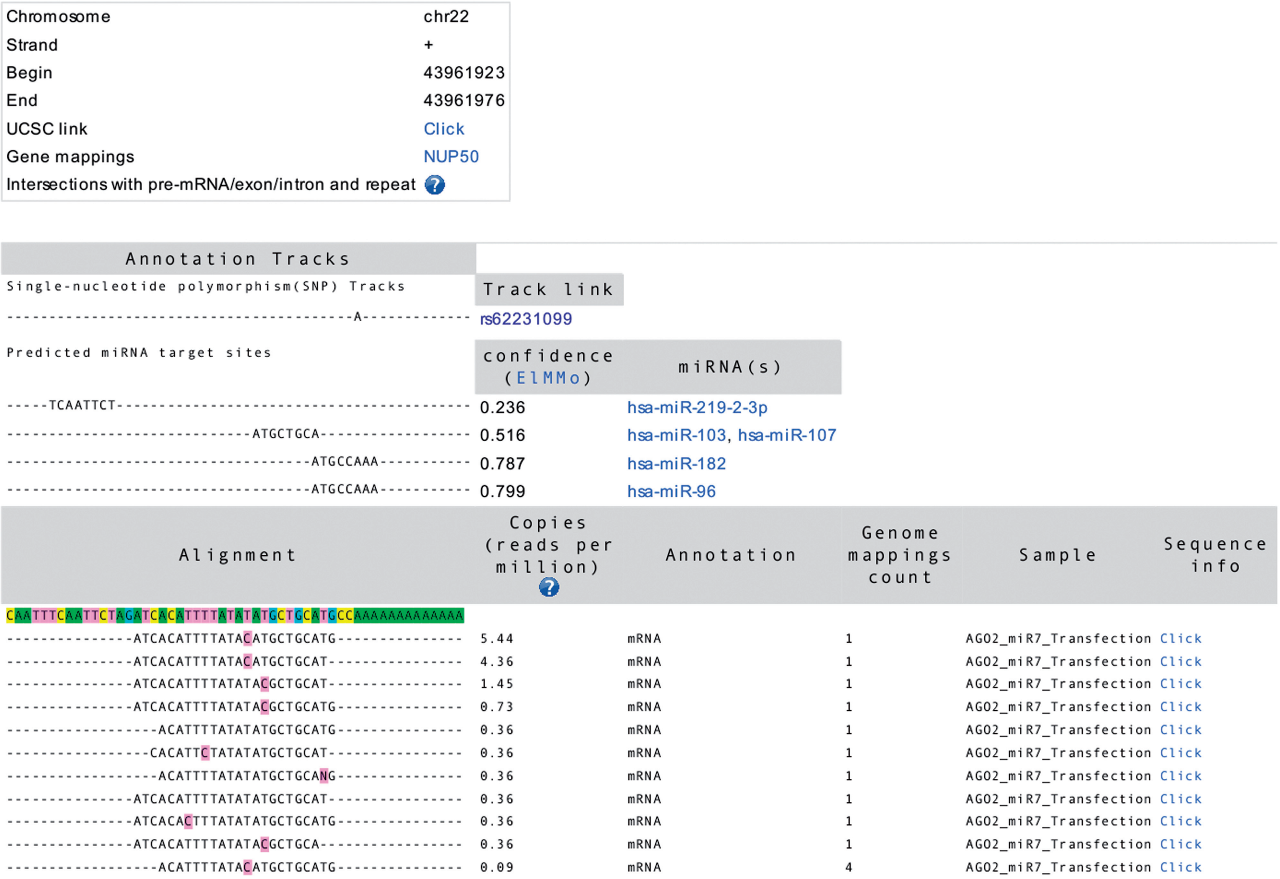


Figure 2. Architecture of the software underlying the CLIPZ analysis environment.

Client. User interactivity is provided by various JavaScript libraries such as:

- Dojo http://www.dojotoolkit.org
- YUI http://developer.yahoo.com/yui
- JQuery http://www.jquery.com

We established a Generic Genome Browser (13) server to generate transcript- or genome-based views of the location of binding sites for one or more proteins in the data set. Communication between the client side and the web server is being established with a Remote Procedure Call (RPC) System that we developed based on the *JsonRPC 2.0* protocol. This is described at http://groups.google.com/group/json-rpc/web/json-rpc-1-2-proposal.



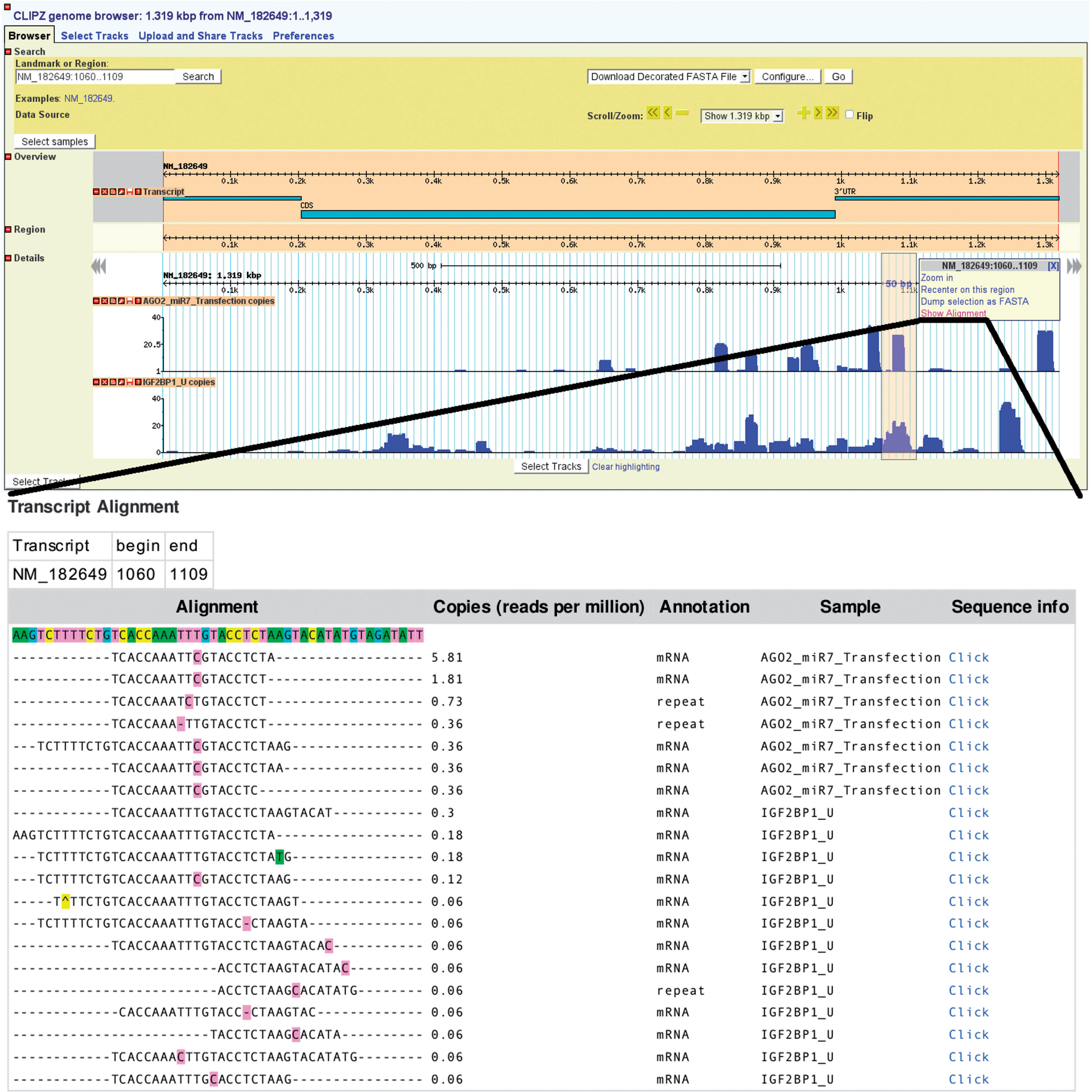
Chromosome	chr22
Strand	+
Begin	43961923
End	43961976
UCSC link	Click
Gene mappings	NUP50
Intersections with pre-mRNA/exon/intron and repeat	?

Annotation Tracks	
Single-nucleotide polymorphism(SNP) Tracks	Track link
`-----------------------------------------A------------`	rs62231099

Predicted miRNA target sites	confidence (ElMMo)	miRNA(s)
`-----TCAATTCT-----------------------------------------`	0.236	hsa-miR-219-2-3p
`-----------------------------ATGCTGCA-----------------`	0.516	hsa-miR-103, hsa-miR-107
`------------------------------------ATGCCAAA----------`	0.787	hsa-miR-182
`------------------------------------ATGCCAAA----------`	0.799	hsa-miR-96

Alignment	Copies (reads per million) ?	Annotation	Genome mappings count	Sample	Sequence info
`CAATTTCAATTCTAGATCACATTTTATATATGCTGCATGCCAAAAAAAAAAAAA`					
`---------------ATCACATTTTATACATGCTGCATG---------------`	5.44	mRNA	1	AGO2_miR7_Transfection	Click
`---------------ATCACATTTTATACATGCTGCAT----------------`	4.36	mRNA	1	AGO2_miR7_Transfection	Click
`---------------ATCACATTTTATATACGCTGCAT----------------`	1.45	mRNA	1	AGO2_miR7_Transfection	Click
`---------------ATCACATTTTATATACGCTGCATG---------------`	0.73	mRNA	1	AGO2_miR7_Transfection	Click
`------------------ACATTTTATATATGCTGCATG---------------`	0.36	mRNA	1	AGO2_miR7_Transfection	Click
`-----------------CACATTCTATATATGCTGCAT----------------`	0.36	mRNA	1	AGO2_miR7_Transfection	Click
`------------------ACATTTTATATATGCTGCANG---------------`	0.36	mRNA	1	AGO2_miR7_Transfection	Click
`---------------ATCACATTTTATATATGCTGCAT----------------`	0.36	mRNA	1	AGO2_miR7_Transfection	Click
`---------------ATCACACTTTATATATGCTGCATG---------------`	0.36	mRNA	1	AGO2_miR7_Transfection	Click
`---------------ATCACATTTTATATACGCTGCA-----------------`	0.36	mRNA	1	AGO2_miR7_Transfection	Click
`------------------ACATTTTATACATGCTGCATG---------------`	0.09	mRNA	4	AGO2_miR7_Transfection	Click

Figure 3. 'Cluster View' of the AGO2/EIF2C PAR-CLIP reads mapping to the nucleoporin 50 kDa (NUP50) gene, also showing single-nucleotide polymorphisms (SNPs) and predicted miRNA binding sites [with their corresponding probabilities given by the ElMMo model (14)] in the neighborhood of the RBP binding site.



Transcript Alignment

Transcript	begin	end
NM_182649	1060	1109

Alignment	Copies (reads per million)	Annotation	Sample	Sequence info
AAGTCTTTTCTGTCACCAAATTTGTACCTCTAAGTACATATGTAGATATT				
------------TCACCAAATTCGTACCTCTA------------------	5.81	mRNA	AGO2_miR7_Transfection	Click
------------TCACCAAATTCGTACCTCT-------------------	1.81	mRNA	AGO2_miR7_Transfection	Click
------------TCACCAAATCTGTACCTCT-------------------	0.73	repeat	AGO2_miR7_Transfection	Click
------------TCACCAAA-TTGTACCTCT-------------------	0.36	repeat	AGO2_miR7_Transfection	Click
---TCTTTTCTGTCACCAAATTCGTACCTCTAAG----------------	0.36	mRNA	AGO2_miR7_Transfection	Click
------------TCACCAAATTCGTACCTCTAA-----------------	0.36	mRNA	AGO2_miR7_Transfection	Click
------------TCACCAAATTCGTACCTC--------------------	0.36	mRNA	AGO2_miR7_Transfection	Click
------------TCACCAAATTTGTACCTCTAAGTACAT-----------	0.3	mRNA	IGF2BP1_U	Click
AAGTCTTTTCTGTCACCAAATTTGTACCTCTA------------------	0.18	mRNA	IGF2BP1_U	Click
---TCTTTTCTGTCACCAAATTTGTACCTCTATG----------------	0.18	mRNA	IGF2BP1_U	Click
---TCTTTTCTGTCACCAAATTCGTACCTCTAAG----------------	0.12	mRNA	IGF2BP1_U	Click
-----T^TTCTGTCACCAAATTTGTACCTCTAAGT---------------	0.06	mRNA	IGF2BP1_U	Click
---TCTTTTCTGTCACCAAATTTGTACC-CTAAGTA--------------	0.06	mRNA	IGF2BP1_U	Click
------------TCACCAAATTTGTACCTCTAAGTACAC-----------	0.06	mRNA	IGF2BP1_U	Click
-------------------------ACCTCTAAGTACATAC---------	0.06	mRNA	IGF2BP1_U	Click
-------------------------ACCTCTAAGCACATATG--------	0.06	repeat	IGF2BP1_U	Click
-------------CACCAAATTTGTACC-CTAAGTAC-------------	0.06	mRNA	IGF2BP1_U	Click
------------------------TACCTCTAAGCACATA----------	0.06	mRNA	IGF2BP1_U	Click
------------TCACCAAACTTGTACCTCTAAGTACATATG--------	0.06	mRNA	IGF2BP1_U	Click
------------TCACCAAATTTGCACCTCTAAG----------------	0.06	mRNA	IGF2BP1_U	Click

Figure 4. 'Transcript View' of the location of AGO2/EIF2C CLIP reads along the transcript with Genbank accession NM_182649, which is the human proliferating cell nuclear antigen (PCNA) transcript variant 2. The coding region (CDS), 5′ and 3′ UTR regions are represented as turquoise-colored boxes. The density of reads from the selected samples (AGO2/EIF2C CLIP performed in miR-7-transfected HEK293 cells and IGF2BP1 CLIP in HEK293 cells) along the entire transcript is shown in blue. The user can select transcript regions and visualize the detailed alignment of reads to these regions.

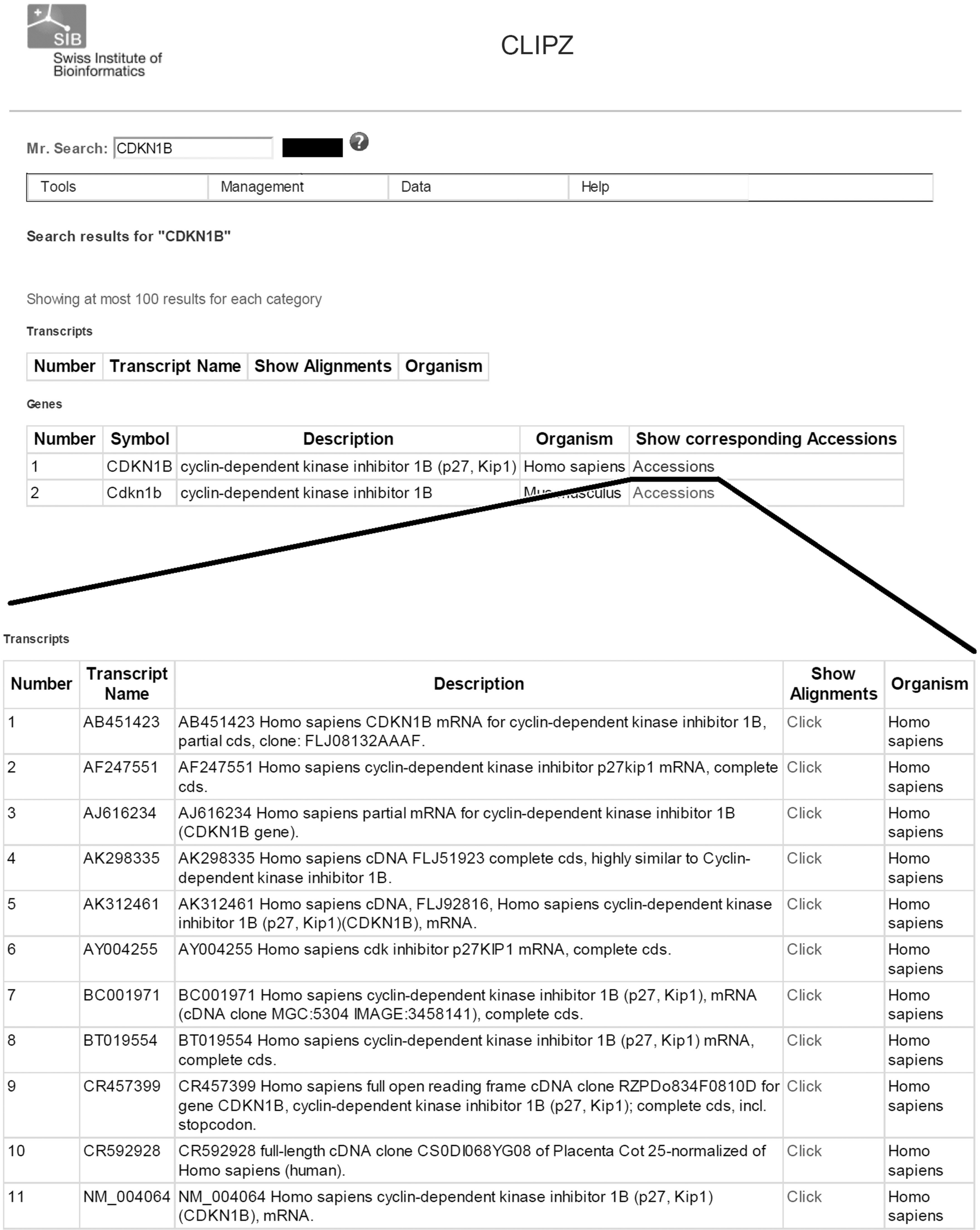
EXAMPLES OF INTERACTIVE ANALYSES

Visualization of clusters of genome- or transcript-based clusters of reads

For each sample in the database, the user can browse the clusters of overlapping sequence reads which in the PAR-CLIP samples typically correspond to individual RBP binding sites. The clusters can be sorted by various criteria including the number of T-to-C mutations in all reads of a cluster, which in the PAR-CLIP experiments is indicative of the affinity of the protein for the RNA. To distinguish cross-link-induced mutations from single nucleotide polymorphisms (SNPs) we incorporated a track that shows the known SNPs, and for identifying the miRNAs that guide the Argonaute to the target RNA, we incorporated a track of predicted miRNA binding sites (14) (see Figure 3).

Transcript and genome browsers

The association of an RBP with a specific site and the downstream effects of this interaction frequently depend



CLIPZ

Mr. Search: CDKN1B [redacted]

Search results for "CDKN1B"

Showing at most 100 results for each category

Transcripts

Number	Transcript Name	Show Alignments	Organism

Genes

Number	Symbol	Description	Organism	Show corresponding Accessions
1	CDKN1B	cyclin-dependent kinase inhibitor 1B (p27, Kip1)	Homo sapiens	Accessions
2	Cdkn1b	cyclin-dependent kinase inhibitor 1B	Mus musculus	Accessions

Transcripts

Number	Transcript Name	Description	Show Alignments	Organism
1	AB451423	AB451423 Homo sapiens CDKN1B mRNA for cyclin-dependent kinase inhibitor 1B, partial cds, clone: FLJ08132AAAF.	Click	Homo sapiens
2	AF247551	AF247551 Homo sapiens cyclin-dependent kinase inhibitor p27kip1 mRNA, complete cds.	Click	Homo sapiens
3	AJ616234	AJ616234 Homo sapiens partial mRNA for cyclin-dependent kinase inhibitor 1B (CDKN1B gene).	Click	Homo sapiens
4	AK298335	AK298335 Homo sapiens cDNA FLJ51923 complete cds, highly similar to Cyclin-dependent kinase inhibitor 1B.	Click	Homo sapiens
5	AK312461	AK312461 Homo sapiens cDNA, FLJ92816, Homo sapiens cyclin-dependent kinase inhibitor 1B (p27, Kip1)(CDKN1B), mRNA.	Click	Homo sapiens
6	AY004255	AY004255 Homo sapiens cdk inhibitor p27KIP1 mRNA, complete cds.	Click	Homo sapiens
7	BC001971	BC001971 Homo sapiens cyclin-dependent kinase inhibitor 1B (p27, Kip1), mRNA (cDNA clone MGC:5304 IMAGE:3458141), complete cds.	Click	Homo sapiens
8	BT019554	BT019554 Homo sapiens cyclin-dependent kinase inhibitor 1B (p27, Kip1) mRNA, complete cds.	Click	Homo sapiens
9	CR457399	CR457399 Homo sapiens full open reading frame cDNA clone RZPDo834F0810D for gene CDKN1B, cyclin-dependent kinase inhibitor 1B (p27, Kip1); complete cds, incl. stopcodon.	Click	Homo sapiens
10	CR592928	CR592928 full-length cDNA clone CS0DI068YG08 of Placenta Cot 25-normalized of Homo sapiens (human).	Click	Homo sapiens
11	NM_004064	NM_004064 Homo sapiens cyclin-dependent kinase inhibitor 1B (p27, Kip1) (CDKN1B), mRNA.	Click	Homo sapiens

Figure 5. The Search tool can be used to retrieve transcripts by Genbank accession number, gene name or symbol. Binding sites in the transcript of interest are then shown through the 'Transcript view'.

on other regulatory elements that are present in close vicinity and recruit other regulatory factors. Through the transcript and genome browsers, one can visualize the position of binding sites within transcripts, as well as the spatial relationship between binding sites determined in different experiments, as shown in Figure 4.

Genome/known sequence super clustering. Many questions arising in the context of analyzing RBP binding sites can be phrased in terms of the spatial relationship between binding sites obtained in different experiments. For example, one would like to know whether experimental results for one protein are reproducible, in which case we expect that the sets of sites obtained in different experiments are largely identical. Alternatively, one may like to find out whether two proteins frequently compete for sites, in which case we would expect that the sites are occupied by one of the proteins in one condition and by the other protein in a different condition. The super-clustering tool enables the user to uncover such relationships.

The visualizations that can be performed are very similar to those described for clusters of a single RBP. But they operate on super-clusters that are built through single-linkage clustering of clusters obtained in different experiments and are either overlapping or at a specified maximum distance from each other. The user may define complex operations between sites obtained in different experiments using logical operators such as (*OR, AND, NOT*).

Search tool. Another common question is whether any binding sites are known for specific transcripts or genes that a user may be studying. To be able to answer this question, we implemented a search tool that allows the user to retrieve from the database a *gene name* or *symbol*, select an accession number associated with it and access the binding site information associated with the transcript in our database (see Figure 5).

miRNA-specific tools. Because the Argonaute/EIF2C proteins that are part of the RNA-induced silencing complex have been a major focus of the CLIP studies performed so far, we integrated in our server a set of tools that enable the user to explore the identity, abundance and predicted targets of the miRNAs present that were isolated in the CLIP samples. These tools have been described extensively in (15).

Motif enrichment tool. Finally, one of the main reasons for performing CLIP studies is to determine the sequence specificity of a protein of interest. How a multi-domain RBP contacts RNAs is a challenging question that most likely requires complex computational as well as experimental analyses. However, to provide some preliminary insights we implemented a tool that identifies sequence motifs (defined as *n*-mers) that are over-represented in an input file (which could contain for instance the most abundant 1000 clusters obtained in an experiment) compared with randomized sequences with the same mono/di-nucleotide composition. We have previously used this tool to show that the motifs that are most over-represented in the clusters from Argonaute/EIF2C PAR-CLIP experiments correspond to the reverse complements of the 5′ end (‘seed’ region) of the most abundant miRNAs in the cell (7).

DISCUSSION

Deciphering the post-transcriptional regulatory code that is implemented by regulatory RNAs and RBPs is a problem of great interest (5–8,16–18). The bottleneck in characterizing RBP binding sites is no longer the availability of an experimental approach, but rather the efficient analysis of the large volumes of data that result from such experiments. Here, we present a software system that we developed to analyze data resulting from CLIP experiments. With this system we constructed a database of RBP binding sites that were determined through CLIP and deep sequencing. Our system provides several views of the data, from the level of sequence reads to that of a whole-genome browser. Transcript regions with the highest abundance in the CLIP data or that exhibit the highest number of cross-linking events can be easily extracted for further analyses. Both the database and the analysis environment can be easily extended. Registered users can expand the database by submitting their own sequence data sets, the repertoire of organisms can be expanded to include additional species for which a genome assembly is available, and the genome assemblies and transcript databases that are used in the analysis pipeline can be updated as necessary. In the future, we will continue to develop the platform in order to accommodate developments in the sequencing technologies. We expect for instance that the increase in sample size and sequence read length will require the use of heuristic algorithms for mapping short reads to the genome. Such algorithms are in fact already available (19–22) and will only require one to write adapter programs to interface these programs with the database that stores the alignments. Thus, CLIPZ can eliminate many bottlenecks in the computational analysis of CLIP data and can form the basis for a repository of binding site data for RNA-binding proteins.

ACKNOWLEDGEMENTS

We are grateful to Lukasz Jaskiewicz, Shivendra Kishore and the other members of the Zavolan group, to Markus Hafner, Manuel Ascano and Markus Landthaler of the Tuschl group for providing input and feedback on individual tools and to Lukas Burger and Jean Hausser for critical comments on the manuscript.

FUNDING

ProDoc program of the Swiss National Science Foundation [Grants PDAMP3_127218 and PDFMP3_123123]; Swiss Institute of Bioinformatics. Funding for open access charge: University of Basel.

Conflict of interest statement. None declared.

REFERENCES

1. Lander,E.S., Linton,L.M., Birren,B., Nusbaum,C., Zody,M.C., Baldwin,J., Devon,K., Dewar,K., Doyle,M., FitzHugh,W.J. *et al.* (2001) Initial sequencing and analysis of the human genome. *Nature*, **409**, 860–921.
2. Izquierdo,J.M. and Valcarcel,J. (2007) Two isoforms of the T-cell intracellular antigen 1 (TIA-l) splicing factor display distinct splicing regulation activities. Control of TIA-l isoform ratio by TIA-l-related protein. *J. Biol. Chem.*, **282**, 19410–19417.
3. Zheng,S., Robertson,T.A. and Varani,G. (2007) A knowledge-based potential function predicts the specificity and relative binding energy of RNA-binding proteins. *FEBS J.*, **274**, 6378–6391.
4. Auweter,S.D., Oberstrass,F.C. and Allain,F.H.-T. (2006) Sequence-specific binding of single-stranded RNA: is there a code for recognition? *Nucleic Acids Res.*, **34**, 4943–4959.
5. Ule,J., Jensen,K.B., Ruggiu,M., Mele,A., Ule,A. and Damell,R.B. (2003) CLIP identifies Nova-regulated RNA networks in the brain. *Science*, **302**, 1212–1215.
6. Chi,S.W., Zang,J.B., Mele,A. and Darnell,R.B. (2009) Argonaute HITS-CLIP decodes microRNA-mRNA interaction maps. *Nature*, **460**, 479–486.

7. Hafner,M., Landthaler,M., Burger,L., Khorshid,M., Hausser,J., Berninger,P., Rothballer,A., Ascano,M., Jungkarnp,A.-C., Munschauer,M. *et al*. (2010) Transcriptome-wide identification of RNA-binding protein and microRNA target sites by PAR-CLIP. *Cell*, **141**, 129–141.
8. König,J., Zamack,K., Rot,G., Curk,T., Kayikci,M., Zupan,B., Turner,D.J., Luscombe,N.M. and Ule,J. (2010) iCLIP reveals the function of hnRNP particles in splicing at individual nucleotide resolution. *Nat. Struct. Mol. Biol.*, **17**, 909–915.
9. Berninger,P., Gaidatzis,D., Nimwegen,E. and van and Zavolan,M. (2008) Computational analysis of small RNA cloning data. *Methods*, **44**, 13–21.
10. Bennett,S. (2004) Solexa Ltd. *Pharmacogenomics.*, **5**, 433–438.
11. Holmes,I. and Durbin,R. (1998) Dynamic programming alignment accuracy. *J. Comput. Biol.*, **5**, 493–504.
12. Wu,T.D. and Watanabe,C.K. (2005) GMAP: a genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics*, **21**, 1859–1875.
13. Stein,L.D., Mungall,C., Shu,S., Caudy,M., Mangone,M., Day,A., Nickerson,E., Stajich,J.E., Harris,T.W., Arva,A. *et al*. (2002) The generic genome browser: a building block for a model organism system database. *Genome Res.*, **12**, 1599–1610.
14. Gaidatzis,D., van Nimwegen,E., Hausser,J. and Zavolan,M. (2007) Inference of miRNA targets using evolutionary conservation and pathway analysis. *BMC Bioinformatics*, **8**, 69.
15. Hausser,J., Berninger,P., Rodak,C., Jantscher,Y., Wirth,S. and Zavolan,M. (2009) MirZ: an integrated microRNA expression atlas and target prediction resource. *Nucleic Acids Res.*, **37**, W266–W272.
16. Keene,J.D. and Lager,P.J. (2005) Post-transcriptional operons and regulons co-ordinating gene expression. *Chromosome Res.*, **13**, 327–337.
17. Galgano,A., Forrer,M., Jaskiewicz,L., Kanitz,A., Zavolan,M. and Gerber,A.P. (2008) Comparative analysis of mRNA targets for human PUF-family proteins suggests extensive interaction with the miRNA regulatory system. *PLoS ONE*, **3**, e3164.
18. Yeo,G.W., Coufal,N.G., Liang,T.Y., Peng,G.E., Fu,X.-D. and Gage,F.H. (2009) An RNA code for the FOX2 splicing regulator revealed by mapping RNA-protein interactions in stem cells. *Nat. Struct. Mol. Biol.*, **16**, 130–137.
19. Li,H. and Durbin,R. (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, **25**, 1754–1760.
20. Li,H. and Durbin,R. (2010) Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics*, **26**, 589–595.
21. Langmead,B., Trapnell,C., Pop,M. and Salzberg,S.L. (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.*, **10**, R25.
22. Wu,T.D. and Nacu,S. (2010) Fast and SNP-tolerant detection of complex variants and splicing in short reads. *Bioinformatics*, **26**, 873–881.